Bilingualism: Language and Cognition

cambridge.org/bil

The use of exemplars differs between native and non-native listening

Annika Nijveld[1,3], Louis ten Bosch[1] and Mirjam Ernestus[1,2]

[1]Centre for Language Studies, Radboud University, Nijmegen, the Netherlands; [2]Max Planck Institute for Psycholinguistics, Nijmegen, the Netherlands and [3]Department of Linguistics, University of Alberta, Edmonton, Canada

Research Article

Cite this article: Nijveld A, ten Bosch L, Ernestus M (2022). The use of exemplars differs between native and non-native listening. *Bilingualism: Language and Cognition* **25**, 841–855. https://doi.org/10.1017/S1366728922000116

Received: 13 May 2019
Revised: 6 October 2021
Accepted: 28 February 2021
First published online: 5 April 2022

Keywords:
speech comprehension; exemplar effects; non-native listening; native listening; processing resources; familiarity; speaker voice; speech reduction

Address for correspondence:
Louis ten Bosch
Erasmusplein 1
6525 HT Nijmegen
E-mail: louis.tenbosch@ru.nl

Abstract

This study compares the role of exemplars in native and non-native listening. Two English identity priming experiments were conducted with native English, Dutch non-native, and Spanish non-native listeners. In Experiment 1, primes and targets were spoken in the same or a different voice. Only the native listeners showed exemplar effects. In Experiment 2, primes and targets had the same or a different degree of vowel reduction. The Dutch, but not the Spanish, listeners were familiar with this reduction pattern from their L1 phonology. In this experiment, exemplar effects only arose for the Spanish listeners. We propose that in these lexical decision experiments the use of exemplars is co-determined by listeners' available processing resources, which is modulated by the familiarity with the variation type from their L1 phonology. The use of exemplars differs between native and non-native listening, suggesting qualitative differences between native and non-native speech comprehension processes.

Introduction

Hybrid models of speech comprehension (e.g., Church & Schacter, 1994; McLennan & Luce, 2005; Goldinger, 2007; Wilder, 2018) distinguish two types of mental representations for the pronunciation of words: abstract representations and exemplars. Abstract representations consist of categorical units (e.g., phonemes), and do not contain fine-phonetic details about each token of a word. Exemplars, in contrast, mentally represent all experienced occurrences of a word in full phonetic detail, which together form a cloud associated with that word.

The present study compares the role of exemplars in native and non-native listening. Many studies in the literature point to a role for exemplars in listeners' NATIVE language (e.g., Goh, 2005; McLennan & Luce, 2005; Palmeri, Goldinger & Pisoni, 1993). The current study investigates whether the use of exemplars differs between native (L1) and non-native (L2) listeners. Importantly, if the use of exemplars differs between native and non-native listening, there are qualitative differences between the relative importance of speech processing mechanisms involved in these two types of listening. Knowledge about the relevance of exemplars in L2 will be relevant for advancing models of L2 speech comprehension (e.g., the 'bilingual interactive activation plus' (BIA+) model developed by Dijkstra & Heuven, 2002), which all ignore the possible role of exemplars.

The representation of words as exemplars for L1 listeners is, despite some mixed results, supported by a range of auditory identity priming studies (e.g., Craik & Kirsner, 1974; Goldinger, 1996; Pufahl & Samuel, 2014). In these experiments, participants recognized repeated words more quickly and/or more accurately if the two tokens of the word ('prime' and 'target') shared perceptual characteristics such as the speaker's voice (the 'match' condition) than when they did not (the 'mismatch' condition – these effects are referred to as 'exemplar effects'). If a prime and corresponding target were both recognized via the same abstract representation, there would be no such difference between matching and mismatching targets. The assumption is therefore that listeners had retained all perceptual details of the prime in memory as an exemplar, which affected the subsequent processing and recognition of the target.

Most of these experiments used variation in SPEAKER VOICE as the basis for the match/mismatch condition, and therefore corroborate that listeners store this type of variation in the form of exemplars. Other variation types for which exemplar effects have been reported include speech rate, emotional tone of voice, fundamental frequency, and the realization of a given single segment (e.g., Church & Schacter, 1994; Janse, 2008; Krestar & McLennan, 2013; Sumner & Samuel, 2005). Native listeners may thus use exemplars with acoustic-phonetic information representing a range of variation types.

While the role of exemplars in native speech comprehension is well-studied, this does not hold for speech comprehension by non-native listeners. There may be a difference between natives and non-natives because exemplar effects have been reported to depend on the

CAMBRIDGE UNIVERSITY PRESS

availability of cognitive resources, which potentially differs between first and second language processing. Work with native listeners has shown that more challenging conditions may lead to either larger (e.g., McLennan & Luce, 2005; Nijveld, ten Bosch & Ernestus, 2015; Nygaard, Burt & Queen, 2000) or smaller exemplar effects (e.g., McLennan, Luce & Charles-Luce, 2003). Non-native listeners may be assumed to have higher processing costs than natives (Segalowitz, 2010). In addition, their cost may depend on their familiarity with the type of variation they hear.

A few studies investigated exemplar effects for speaker voice for non-native listeners, which is a variation type known to native listeners of all languages (e.g., Winters, Lichtman & Weber, 2013). For instance, Drozdova, van Hout, and Scharenborg (2019) tested exemplar effects of speaker voice in noise and clean speech for Dutch non-native listeners of English. Participants engaged in a word identification task and in an old-new judgment task. Exemplar effects arose in the old-new judgment task only, which does not require word comprehension, and the effects were larger for non-native listeners with higher proficiency levels. The intermediating effect of noise was less clear, as the two dependent variables showed different results: exemplar effects in accuracy were larger in noise, while in the reaction times, the effects were larger in clean speech. Like the study by Winters et al., this study suggests that non-native listeners may show exemplar effects for variation of speaker voice in an old-new judgment task.

Trofimovich (2005) also reported exemplar effects for speaker voice, but in a very different type of experiment. English listeners with self-reported 'low-intermediate proficiency' in Spanish were tested on familiar words in English (i.e., L1 stimuli) and Spanish (i.e., L2 stimuli). Listeners had to repeat out loud words that were presented auditorily to them; these prompts were presented in either the same or a different voice as during a familiarization phase earlier in the experiment. For the Spanish but not the English words, listeners were quicker to start producing the words played in the same voice than in a different voice. Hence, exemplar effects arose in a non-native word repetition task. It is unclear whether there was a difference between L1 and L2 processing because the difference in the size of the exemplar effects was not statistically assessed. This study confirms Drozdova et al. (2019) and Winters et al. (2013) that also L2 listeners may store information about speaker voice and use it for subsequent word processing, but it does not reliably show differences in exemplar effects between L1 and L2 processing.

Morano, ten Bosch, and Ernestus (2019) studied exemplar effects for reduced pronunciation variants of words. Dutch L2 listeners with intermediate proficiency levels in French engaged in a French long-term lexical decision priming task. The two tokens of a word (prime and target) matched or mismatched in the VOICING of the vowel in the word's initial syllable, a variation that is less common in Dutch than in French. Morano and colleagues conducted three versions of their experiment that differed in the word tokens that were used. In versions AA and BB, primes and targets in the match condition were identical tokens taken from either token set A or B (two sets of recordings of the same words made under highly similar circumstances), respectively. In version AB, primes and targets were always different tokens (i.e., even in the match condition). Exemplar effects arose in experiment version AA only. The emergence of exemplar effects in that study was thus highly restricted, and seemed to depend on specific stimulus characteristics. Morano et al. did not compare L2 listeners with L1 listeners.

In sum, exemplar effects have been reported for non-native listeners but it is still unclear whether exemplar effects differ between native and non-native listeners when they are involved in a task that requires lexical access (rather than in old-new judgments or word repetition). Similarly, it is unknown whether potential differences in exemplar effects between native and non-native listeners may depend on their familiarity with the variation given the listener's native language. Listener's familiarity is likely to affect processing load, which may modulate the emergence of exemplar effects.

We address these two questions in the present study by testing native and non-native listeners (English native listeners, and Dutch and Spanish non-native listeners of English) on a task that requires lexical access, a lexical decision task, conducted in English. A lexical decision task places different cognitive demands on native and non-native listeners: while it is quite an easy task for native listeners, this is not true for non-native listeners. Non-native listeners' weaker and less well-specified lexical representations (e.g., Cook, Pandža, Lancaster & Gor, 2016) introduce a considerable amount of uncertainty to the task, which makes it harder for them to perform well in a lexical decision experiment. If processing load affects the emergence of exemplar effects, we therefore expect differences between native and non-native listeners with this task.

We carried out two experiments with two different variation patterns. In Experiment 1, as in the majority of the literature, we used speaker voice as the basis for the match/mismatch condition: repeated words were presented in the same (male – male) or a different (female – male) voice. This variation is L1-unspecific. All listeners have extensive experience processing this variation type from their L1. The experimental words were high-frequency British English words. This experiment thus investigates whether exemplar effects differ or are equivalent between native and non-native listeners for a common variation type. If there are differences, they likely result from differences in processing load.

In Experiment 2, we used variation stemming from speech reduction. The way speakers reduce words is partly language-specific. Native listeners comprehend reduced pronunciation variants (occurring in running speech) with remarkable ease. In contrast, non-native listeners may experience problems understanding reduced speech (e.g., Felker, Ernestus & Broersma, 2019; ten Bosch, Giezenaar, Boves & Ernestus, 2016), probably because their exposure to L2 is often primarily classroom based, where reduction is rare (e.g., Edstrom, 2006; Jones & Ono, 2000; McCarthy & Carter, 1995). However, non-native listeners can be familiar with a given speech reduction pattern through their L1, and overlap between L1 and L2 speech reduction patterns may aid the perception of reduced pronunciation variants in L2 (e.g., Mitterer & Tuinman, 2012).

Reduced tokens in our experiment were overall shorter than the unreduced tokens. More importantly, they had highly shortened vowels in their initial, unstressed syllables (e.g., *balloon* with a very short schwa). This vowel reduction pattern is highly frequent in English (e.g., Dalby, 1986; Shockey, 2003) and Dutch (e.g., Ernestus, 2000). Vowel reduction mostly occurs in words embedded in sentences rather than in words presented in isolation; in Experiment 2, words are presented in isolation. For the English and the Dutch listeners, Experiment 2 may therefore require more processing costs than Experiment 1, which may affect the emergence of exemplar effects.

https://doi.org/10.1017/S1366728922000116 Published online by Cambridge University Press

For the Spanish listeners, the processing costs can be expected to be highest, because vowels in Spanish are hardly reduced (e.g., Torreira & Ernestus, 2011). If processing load affects the emergence of exemplar effects, the Spanish may thus pattern with the Dutch listeners in the presence versus absence of exemplar effects in Experiment 1, while they may not in Experiment 2. The combination of Experiments 1 and 2 is thus informative about the role of the listener's familiarity with a variation pattern on the emergence of exemplar effects in L2 listening.

Experiment 1

Method

Participants

One-hundred and thirteen listeners took part in the experiment (which excludes two listeners whose data we could not use due to technical issues). Of these listeners, 40 were native listeners of English (mean age: 21 years; 6 left-handed; 11 male), 40 were native listeners of Dutch (mean age: 21 years; 6 left-handed; 6 male), and 33 were native listeners of Spanish (mean age: 22 years; 5 left-handed; 21 male). All listeners were highly educated, reported no hearing disorders, gave their informed written consents, and were paid for their participation.

We assessed our non-native listeners' English proficiencies with the LexTALE proficiency task (Lemhöfer & Broersma, 2012; see the Procedure section of Experiment 1 below), on which the Dutch listeners obtained an average score of 74% (*Standard Deviation,* henceforth *SD,* = 11%), while the Spanish listeners were at 67% on average (*SD* = 9%). While both of these averages fall within CEFR level B2 ('upper intermediate' proficiency: 60–80%), a linear regression model showed that the difference between the two groups was statistically significant ($\beta_{Spanish}$ = −6.66, *SE* = 2.33, *t* = -2.86, *p* = .0055). In line with this, self-rated English proficiency (on a 1 - 6 scale; see the Procedure section of Experiment 1 below) was at 4.8 on average for the Dutch listeners (*SD* = 1.0) and at 4.3 for the Spanish listeners (*SD* = 0.9). In order to see whether these differences in English proficiencies may explain differences in exemplar effects between the groups of non-native listeners over and above differences in their native languages, we not only analyzed the full set of data, but we also performed analyses on a subset of the data with non-native listeners who were very comparable in their English proficiencies. This subset included 29 Dutch listeners (average LexTALE score: 69%, *SD* = 7%), 29 Spanish listeners (average LexTALE score: 69%, *SD* = 8%), and a random selection of 29 native English listeners. The mean LexTALE scores of these Dutch and Spanish listeners were not statistically different ($\beta_{Spanish}$ = -0.34, *SE* = 1.94, *t* = -0.2, *p* = .86). Importantly, the statistical analyses on the full data set and on this subset yield similar results with respect to our predictors of interest (i.e., regarding exemplar effects), which indicates that the critical findings reported below also hold for Dutch and Spanish non-native listeners with similar general proficiency levels.

Materials

The experiment contained 43 bi- and trisyllabic real English nouns with stress on the second syllable (of which 30 served as experimental words, ten as real word distractor foils, and three as real word practice items) and an equal number of counterpart pseudo words (listed in Table A1-1 in Appendix 1). Some of the real nouns could be interpreted as verbs as well (e.g., *collapse*). We derived a pseudo word from each real word by keeping the initial syllable, and by altering up to three phonemes in the following syllables through substitution or deletion (e.g., we derived pseudo word *ballee* from real word *balloon*). This procedure resulted in pseudo words that were roughly equally long as and phonologically similar to the real words. While the pseudo words were non-existing, they obeyed English phonotactic constraints. We checked whether participants could readily guess from which real words the pseudo words were derived, by asking four Dutch native listeners (from the target population but not participating in Experiment 1 or 2) to indicate whether the words strongly reminded them of particular real English words. If so, we altered additional phonemes until this was no longer the case.

In the first part of the experiment, our 30 experimental real words occurred as primes, in addition to their 30 counterpart pseudo words. The second part of the experiment contained repeats from the first part: the 30 experimental real words appeared as targets, in addition to their counterpart pseudo words. Additionally, the second part of the experiment contained 20 new distractor foils (i.e., these were not repeats from the first part), consisting of ten real words and their ten counterpart pseudo words. These distractor foils served to lead participants' attention away from stimulus repetition in the experiment to some extent (such that 75%, instead of 100%, of the stimuli in the second part of the experiment were repeats from the first part). Both parts of the experiment started with the same six practice trials, which consisted of three real words and their three counterpart pseudo words.

The experimental real words had an average log-transformed frequency of occurrence of 4.48 (*SD* = 1.97); real word distractor foils of 4.21 (*SD* = 2.58), and real word practice items of 2.54 (*SD* = 2.26; these are log-transformed raw counts from version 1.0 of the British National Corpus, 1995, a corpus containing 100 million word tokens.). These relatively high frequencies make it likely that the non-native listeners are familiar with these words. A Welch two-sample t-test (an adaptation of the regular t-test which can handle samples with unequal variances and which adjusts the number of degrees of freedom accordingly) showed that there was no significant difference in the frequency of occurrence between the thirty experimental real words and the ten real word distractor foils (*t* (12.67) = -0.29, *p* = .77; the degree of freedom (df) amounts to 12.67 as a result of the Satterthwaite-Welch adjustment of the df).

Of the 30 experimental real words, 19 can be considered cognates with Dutch, 23 with Spanish, and 17 with both Dutch and Spanish (see Supplementary Materials for a list of all cognates). We included cognates in the experiment because it was not possible to find sufficient English nouns meeting our criteria that were not cognates (i.e., having a relatively high frequencies of occurrence so that they are likely to be known by our participants, starting with an unstressed syllable, followed by a stressed syllable). The influence of words' cognate status on the occurrence of exemplar effects for non-native listeners is not within the scope of the present study, and our study is not designed to test for such effects. Nevertheless, we wished to explore whether our experimental words' cognate status for the Dutch and Spanish listeners influenced the occurrence of exemplar effects. As such, we carried out additional statistical analyses on the separate data from the Dutch and Spanish listeners where we tested for interactions between exemplar effects and cognate status. These analyses (reported in the Supplementary Materials)

https://doi.org/10.1017/S1366728922000116 Published online by Cambridge University Press

do not suggest that exemplar effects are modulated by the word's cognate status.

We recorded all real and pseudo words (i.e., experimental real words, real word distractor foils, real word practice items, and all of their counterpart pseudo words) with a male native speaker of British English, and we also recorded the experimental real words and their counterpart pseudo words with a female native speaker of British English. The speakers read the real and pseudo words from paper in a sound-attenuating booth, and they were recorded with a Sennheiser ME 64 microphone and Adobe Audition 1.5 recording software at a sampling rate of 44.1 kHz at 2 bytes/sample. We recorded multiple tokens for each real word and pseudo word with each speaker. Editing of the individual sound files for each word (e.g., cutting the individual words and pseudo words from the long audio file and amplitude equalizing) was done with Praat software (Boersma & Weenink, 2018).

Depending on whether the real or pseudo word was to occur only as target (i.e., the real word distractor foils and their counterpart pseudo words) or as both prime and target (i.e., the experimental real words, real word practice items, and each of their counterpart pseudo words) in the experiment, we selected the one or two best sounding tokens from the male speaker's recordings. From the female speaker's recordings, we only selected the best sounding token for each experimental real word and its pseudo word counterpart (because the real and pseudo words produced by the female speaker were only to appear as primes, see also below). The prime and target tokens of the experimental real words produced by the male speaker had average durations of 551 ms (SD = 74 ms, range 450–696 ms) and 571 ms (SD = 70 ms; range 481–740 ms), respectively, while the average duration of the prime tokens of the experimental real words produced by the female speaker was 642 ms (*SD* = 62, range: 551–831 ms). The primes produced by the male and female speakers thus differed considerably in duration.

We presented the prime (in the first part of the experiment) and the target (in the second part) tokens of a word in either the same or a different voice (i.e., in the match or the mismatch conditions, respectively). A match meant that both prime and target were uttered by the male speaker, while a mismatch meant that the prime was uttered by the female speaker and the target by the male speaker (also see Table Supp-1 in Supplementary Materials).

For the first part of the experiment, we created four lists, in each of which half of the stimuli were uttered by the female speaker and half by the male speaker. These lists contained the experimental real words and their counterpart pseudo words. The lists had different pseudo-randomized stimulus orders, and differed in which primes were produced by which speaker. Maximally three real or pseudo words followed each other. For each of the four lists, we created a mirror list in which we replaced the tokens produced by one speaker by tokens produced by the other speaker.

For the second part of the experiment, we again created four pseudo-randomized lists, in which the maximal consecutive number of real or pseudo words was also three. These lists contained new tokens of the experimental words and their counterpart pseudo words (i.e., repetitions from part 1) as well as tokens for the distractor foil real and pseudo words. All stimuli in the second part were produced by the male speaker. We paired these four lists with the four lists and their mirror lists for part 1 (leading to a total of eight pairings); every participant heard one pair of lists. All lists started with six practice trials that were produced by the male speaker, a different set of tokens per part. The order of presentation of the practice items also differed per part, but was the same for all participants.

Procedure

Participants were tested individually in a sound-attenuating booth (the English listeners at Cambridge University in the U.K., the Dutch listeners at the Max Planck Institute for Psycholinguistics in Nijmegen, the Netherlands, and the Spanish listeners at the Escuela Técnica Superior de Ingenieros de Telecomunicación of the Universidad Politécnica de Madrid in Spain). We presented stimuli via closed headphones at a comfortable listening level using E-prime 2.0 software (Psychology Software Tools, Pittsburgh, PA). We instructed participants to decide as quickly and accurately as possible whether the stimulus they heard was a real English word or not. Participants responded by pressing *yes* with their dominant hand on a key board (key *m* for right-handers, *z* for left-handers) or *no* with their non-dominant hand (*z* for right-handers, *m* for left-handers). The two parts of the experiment directly followed each other.

Each trial started with a blank screen, visible for 300 ms. A fixation asterisk then appeared in the middle of the screen for 250 ms, followed by the auditory presentation of the stimulus (blank screen). The next trial started after the response, or in case of no response after three seconds from word onset. We recorded participants' accuracies and response times (RTs) from word onset. This experiment took approximately 20 minutes.

For the two groups of non-native listeners, the main experiment was followed by the LexTALE task (Lemhöfer & Broersma, 2012) and a questionnaire. LexTALE is an unspeeded visual lexical decision task testing English proficiency. Although it tests for vocabulary knowledge, its results have been shown to correlate substantially with general proficiency. The stimuli (three practice items, 40 real English words, and 20 pseudo words) were presented one by one in a pseudo-randomized order on a computer screen in a black font (Arial Unicode MS, point size 18) in the middle of a white background with E-prime 2.0 software. Participants responded via the keyboard, by pressing *yes* with their dominant hand (key *m* for right-handed, *z* for left-handed) and *no* with their non-dominant hand (*z* for right-handed, *m* for left-handed). The next trial appeared on the screen upon the response key press. After the LexTALE task, participants filled out a language background questionnaire, in which they described their experience with English, and in which they self-rated their proficiency levels (on a scale ranging from 1, indicating a very low level, to 6, indicating a near native proficiency level). The LexTALE task and the questionnaire each took approximately five minutes.

Analyses

We statistically analyzed responses to the target occurrences of our experimental real words (henceforth: targets). Prior to the analyses, we excluded listeners and targets whose error rates were separated 2.5 *SD* or more from the overall average error rates for listeners and targets (pooled over listener groups), respectively. We also discarded targets whose primes did not receive correct responses: if primes were not responded to correctly, listeners may not have paid attention to them and/or understood them.

For the analysis of the RTs, we log-transformed the RTs to reduce skewness in the data (as recommended by e.g., Baayen, 2008; Lo & Andrews, 2015). We then excluded targets with

https://doi.org/10.1017/S1366728922000116 Published online by Cambridge University Press

incorrect responses, and targets with RTs deviating 2.5 *SD* from the grand mean. In the Results and discussion section below, we report the overall percentages of excluded trials after each data cleaning step, and indicate the number of excluded targets per listener group. After running the statistical model on the log RTs, we discarded RTs that deviated more than 2.5 residual standard error from the predicted RTs because these were considered outliers, after which we refitted the model.

We analyzed the accuracies of the responses to the targets in the auditory lexical decision task with logistic mixed effects regression models with the binomial link function, and we analyzed the log RTs to targets with mixed effects regression models using the lme4 package (version 1.1.13; Bates, Maechler, Bolker & Walker, 2015). Both types of analyses were conducted in R statistical software (version 3.4.1; R Core Team, 2017).

In both the log RT and accuracy analyses, we tested the influence of our predictors of interest Speaker match (reflecting whether a prime and target were uttered by the same or a different speaker) and Listener group, and the interaction between these. Speaker match was contrast coded, with the level of speaker mismatch included in the intercept. Listener group was Helmert-coded (e.g., Fox & Monette, 2002) to assess within the same statistical model if a) the two groups of non-native listeners differed from the native listeners (Contrast 1), and b) if the two non-native groups differed amongst each other (Contrast 2). More specifically, Contrast 1 compares native (0.666) to non-native listeners (both -0.333), and Contrast 2 compares Dutch (0.5) to Spanish (-0.5) listeners. The coefficients we find in the statistical models should be multiplied with these numbers in order to determine these levels' exact effects.

To capture additional variance in our data, we tested the effects of a number of control predictors that have been shown to be relevant in similar experiments (e.g., Hanique, Aalders & Ernestus, 2013; Morano et al., 2019): log-transformed Word frequency as obtained from the British National Corpus (version 1.0, 1995), Trial number, and Lag in terms of the number of trials between prime and target. Specific to the analysis of the log RTs, we added the control predictors log-transformed reaction times to the prime (Log RT prime) and to the preceding trial (Log RT preceding trial), and the log-transformed Target duration.

We used Word and Listener as crossed random effects in the analyses of the accuracy and the log RTs, and tested for random slopes for the predictors of interest and their interactions (one random slope under listener: Speaker match; two random slopes under word: Listener Group, Speaker match, and Speaker match x Listener Group). We only included random slopes if they did not lead to convergence errors or singular fits, if they were not highly correlated with the random intercept (criterion: $r > .80$), and if their addition significantly improved the model fit. We determined the latter by comparing a model with the random effect to the identical model without this effect, using Chi-square tests performed with R's anova() function. We did not test for random slopes of our control predictors (as Barr, Levy, Scheepers & Tily, 2013 would recommend) for three reasons: first, we had no experimental hypotheses about those; second, doing so increases the chances of overfitting the models to this particular dataset (which decreases the generalizability of our findings); and third, doing so increases the chances of model convergence failures (cf. Bates, Kliegl, Vasishth & Baayen, 2015).

Below, we report the models only including statistically significant effects and interactions as well as simple effects of predictors appearing in statistically significant interactions. We determined significance by examining whether the absolute z (in case of the accuracy analysis) or t (in case of the log RT analysis) values of the effects' coefficients in absolute value exceeded 1.96, which implies $p < .05$ for this type of data with many observations that together approach a normal distribution. We did not include fixed effects if they led to model convergence issues. When random slopes were not included in our final models, Appendix 2 details the reasons for not including them. In Appendix 2, we also list the 'full' statistical models in which we included all fixed effects we tested (whether or not they were significant) as long as the effects' inclusion did not lead to model convergence issues.

The final models that we report in the result sections were thus obtained from the full models after systematically removing the insignificant predictors – predictors were kept when they appeared in significant interactions. Where we report statistical information of statistically non-significant effects in the running text, they are taken from the full models.

As described above, all targets were uttered by the male speaker. One may wonder whether this set-up may have alerted listeners to disregard what they previously heard from the female speaker. If so, there should be no priming on targets preceded by primes produced by the female speaker. To verify whether these targets were primed, we statistically compared targets preceded by primes produced by the female speaker to primes produced by the male speaker in an additional analysis (all stimuli analyzed here were thus produced by the male speaker). If priming occurred, responses to the targets following primes produced by the female speaker should be more accurate and/or faster than responses to *primes* produced by the male speaker.

Results and discussion

According to the outlier criteria on error rates described above, we excluded two Spanish and one Dutch non-native listener as well as the target *saloon* from the analyses. In addition, we excluded trials with targets whose primes were responded to incorrectly (8% of the data; comprising 34 targets responded to by the native English listeners, 86 by the Dutch listeners, and 122 by the Spanish listeners). Listeners then, on average, made 3% of errors on the targets (overall *SD* 3%; native English listeners: 2%; Dutch listeners: 3%; Spanish listeners: 4%; $SD = 4\%$, 3%, 3%, respectively).

We first investigated whether our listeners took the primes produced by the female speaker into account when listening to the targets, which were all produced by the male speaker. We found that the primes produced by the female speaker primed the targets: responses to targets following primes produced by the female speaker had a significantly higher likelihood of receiving correct responses than responses to primes produced by the male speaker had ($\beta_{target} = 0.82$, $SE = 0.16$, $z = 5.10$, $p = 3.0\text{e-}7$). In a context of targets uttered by the male speaker only, listeners did thus not appear to disregard primes produced by the female speaker.

We then investigated which factors could predict response accuracy for all targets. Table A2-1 shows the full model, listing all predictors that we could test without obtaining convergence errors. No predictor was statistically significant. We found no effect of Listener group, showing there was no statistically significant difference between the English listeners on the one hand and the Dutch and Spanish listeners on the other hand (i.e., Contrast

https://doi.org/10.1017/S1366728922000116 Published online by Cambridge University Press

1; $\beta = 0.32$, $SE = 0.33$, $z = 0.97$, $p = .33$), or between the Dutch and the Spanish listeners (i.e., Contrast 2; $\beta = 0.27$, $SE = 0.38$, $z = 0.71$, $p = .48$). We also did not find a simple effect of Speaker match ($\beta_{\text{speaker match}} = 0.38$; $SE = 0.22$, $z = 1.69$, $p = .09$), nor a statistically significant interaction between Speaker match and Listener group ($\beta_{\text{speaker match x contrast 1}} = 0.26$; $SE = 0.48$, $z = 0.54$, $p = .59$; $\beta_{\text{speaker match x contrast 2}} = 0.10$; $SE = 0.53$, $z = 0.20$, $p = .84$).

For the log RT analysis, we excluded targets with incorrect responses (this yielded, together with the targets whose primes received incorrect responses, an exclusion of 4% of the data; 62 of the targets responded to by English native listeners, 118 by the Dutch listeners, and 153 by the Spanish listeners). We then removed outliers relative to the grand RT mean (3% of the data; 20 targets by the native English listeners, 27 by the Dutch listeners, and 36 by the Spanish listeners). Remaining RTs ranged from 558 ms to 1656 ms from word onset, and were 936 ms on average ($SD = 186$ ms).

Again, we first examined whether primes produced by the female speaker primed their targets. As in accuracy, we found evidence for this: responses to targets following primes produced by the female speaker were significantly faster (947 ms on average) than to primes produced by the male speaker (1033 ms; $\beta_{\text{target}} = -0.08$, $SE = 0.005$, $t = -14.9$, p = 1.0 e-16). This finding again indicates that listeners did not disregard primes produced by the female speaker.

We then built a model predicting the log RTs to the targets. We did not include Listener group as random slope by Word because its inclusion led to model convergence issues. Speaker match was not included as random slope by Word nor by Listener because this slope was highly correlated with each of those intercepts ($r = 1.0$).

Our final statistical model (see Table 1) shows effects of the control predictors Log RT previous, Log RT prime, and Log target duration, indicating faster responses to targets whose primes or preceding trials received quicker responses and to shorter targets. We also observed an effect of Contrast 1 of Listener group, showing that the native English listeners responded significantly more quickly than the non-native listeners did.

Of relevance to our research question, we observed a significant interaction between Speaker match and Listener group. The statistical model shows that the effect of Speaker match differed between the native listeners and non-native listeners (Contrast 1), but not between the Dutch and Spanish listeners (Contrast 2; see Figure 1). We ran statistical models without the simple and interaction effects of Listener group on the data split according to Contrast 1 to interpret the difference between the native and the non-native listeners. The model for the native listeners showed a significant effect of Speaker match ($\beta_{\text{speaker match}} = -0.018$, $SE = 0.0083$, $t = -2.13$, p = 0.033), whereas in the model for the non-native listeners Speaker match was not statistically significant ($\beta_{\text{speaker match}} = -0.0039$, $SE = 0.052$, $t = -0.74$, p = 0.46).

Appendix 2 Table A2-2 shows the model in which all predictors are included, including ones that were not statistically significant and thus not included in our final model.

Experiment 2

Method

Participants

A total of 114 listeners (who did not participate in Experiment 1) took part in Experiment 2 (which excludes seven listeners whose data we could not use due to technical issues, two listeners reported hearing issues, and two listeners who reported a deviant native language compared to the target sample). Thirty-four were native English listeners (13 male, five left-handers, mean age: 22 years), 40 were native Dutch listeners (11 male, two left-handed, mean age: 20 years, mean LexTALE score: 73%, $SD = 12\%$; mean self-assessed listening proficiency: 4.7, $SD = 0.7$), and 40 were native Spanish listeners (24 male, all right-handed, mean age: 22 years; mean LexTALE score: 67%, $SD = 10\%$; mean self-assessed listening proficiency: 3.8, $SD = 1.1$). The two non-natives groups' LexTALE scores fall within the range of CEFR level B2.

As for Experiment 1, the LexTALE scores of the Dutch listeners were significantly higher than those of the Spanish listeners ($\beta_{\text{Spanish}} = -6.3$, $SE = 2.4$, $t = -2.6$, $p = 0.0093$). We therefore again conducted analyses on a subset of the data comprising non-native listeners with the most similar LexTALE scores as well as the native listeners. This subset included 33 Dutch listeners (with an average LexTALE of score 69%, $SD = 9\%$), 33 Spanish listeners (average LexTALE score: 69%, $SD = 9\%$), and all 34 English native listeners. The mean LexTALE scores of the Dutch and Spanish listeners in this subset were not statistically different ($\beta_{\text{Spanish}} = -0.19$, $SE = 2.18$, $t = -0.09$; $p = .93$). Importantly, the analyses performed on the subset yielded the same patterns as the analyses on the full data set with respect to our predictors of interest. As in Experiment 1, all results reported below thus also hold for Dutch and Spanish non-native listeners with similar proficiencies in English.

Materials

We used the same word types as in Experiment 1. The two experiments differ in the possible difference between the two occurrences of a word type. In Experiment 1, stimulus repetitions could be in the same or a different voice (male - male or female - male). In Experiment 2, stimuli were reduced or unreduced pronunciation variants, all produced by the same speaker. While the primes could be unreduced or reduced, the targets were all reduced. The speaker was the male speaker from Experiment 1. The reduced pronunciation variants are characterized by shorter overall durations and shortened segments typical of a casual speech style; in particular, the vowel in the initial syllable was substantially shorter. As an example, a reduced and an unreduced token of the experimental real word *cassette* are shown in Figure 2.

We re-used stimuli produced by the male speaker from Experiment 1 as unreduced stimuli in Experiment 2. For the reduced stimuli, we made new recordings with the same speaker and the same recording equipment. The speaker was instructed to produce tokens as if they were produced in casual speech. Because the pseudo words had the same initial syllables as the real words, schwa reduction was also possible in the pseudo words. We selected the two best sounding tokens for to-be-repeated stimuli (i.e., experimental words, practice items and their respective counterpart pseudo words), and the single best sounding tokens for stimuli that were not to be repeated (i.e., the real word distractor foils and their counterpart pseudo words).

Unreduced tokens of the experimental real words had an average duration of 628 ms ($SD = 75$ ms); the reduced tokens had an average duration of 528 ms ($SD = 68$ ms). The durations of the reduced and unreduced primes and reduced targets are illustrated in Figure 3. Even though all reduced word tokens had a small portion of the initial vowel left, in these tokens it was perceptually close to absent.

https://doi.org/10.1017/S1366728922000116 Published online by Cambridge University Press

Table 1. Final statistical model for log RTs of correct responses to targets in Experiment 1. The intercept represents *Speaker mismatch*. *SE* stands for Standard Error. Contrast 1 compares native (0.666) to non-native listeners (both -0.333), and Contrast 2 compares Dutch (0.5) to Spanish (-0.5) listeners.

Fixed effects	*B*	*SE*	*t*	*p*
Intercept	4.21	0.48	8.73	1 e-16
Speaker match (match)	−0.0036	0.052	−0.69	0.49
Log RT previous	0.063	0.011	5.58	1.9 e-9
Log RT prime	0.180	0.014	12.52	1 e-16
Log target duration	0.147	0.074	2.01	0.045
Contrast 1 (native vs. non-native)	−0.066	0.017	−4.19	4.1 e-5
Contrast 2 (Dutch vs. Spanish)	−0.0047	0.019	−0.26	0.80
Speaker match (match) x Contrast 1	−0.022	0.011	−2.12	0.034
Speaker match (match) x Contrast 2	−0.012	0.013	−0.89	−0.37
Random effects			*SD*	
Word	Intercept		0.045	
Listener	Intercept		0.070	
Residual			0.132	

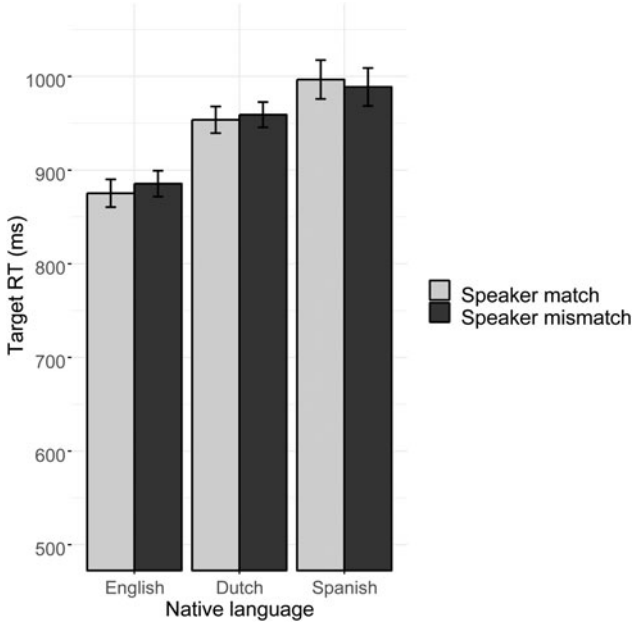


Fig. 1. (Raw, i.e., non log-transformed) RTs of analyzed correct responses to targets in Experiment 1, split according to listeners' native languages, and speaker match condition. Error bars represent 95% confidence intervals.

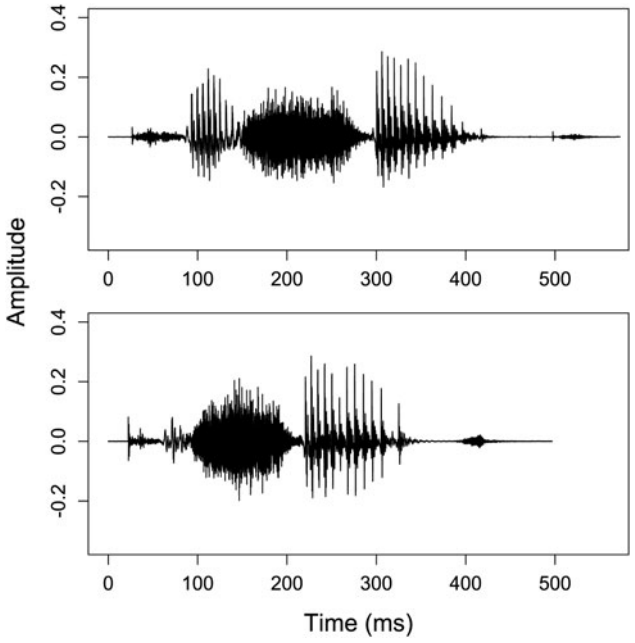


Fig. 2. Examples of stimuli: an unreduced (top) and a reduced (bottom) token of the experimental real word *cassette* /kəˈsɛt/. The figure shows a substantial difference in the two tokens' overall duration as well as in the realization of the vowels in the two tokens' initial syllables.

As for Experiment 1, we investigated for Experiment 2 whether the cognates may behave differently from the non-cognates. Again, we carried out additional statistical analyses on the separate data of the Dutch and Spanish listeners and tested for interactions between exemplar effects and cognate status. These analyses (reported in the Supplementary Materials) do not suggest that exemplar effects are modulated by the word's cognate status.

The experimental lists were identical to the ones used in Experiment 1, except that we adapted them to the new manipulation. We replaced all male speaker's tokens by reduced tokens, and replaced all female speaker's tokens by unreduced tokens.

Procedure and analyses

The procedure and analyses were identical to the ones in Experiment 1, except that the predictor of interest Speaker match was replaced by Variant match.

Results and discussion

According to our 2.5 *SD* outlier exclusion criterion for errors, we excluded three Spanish non-native listeners, and, as in Experiment 1, we discarded the target *saloon* from the analyses. Prior to all of

https://doi.org/10.1017/S1366728922000116 Published online by Cambridge University Press

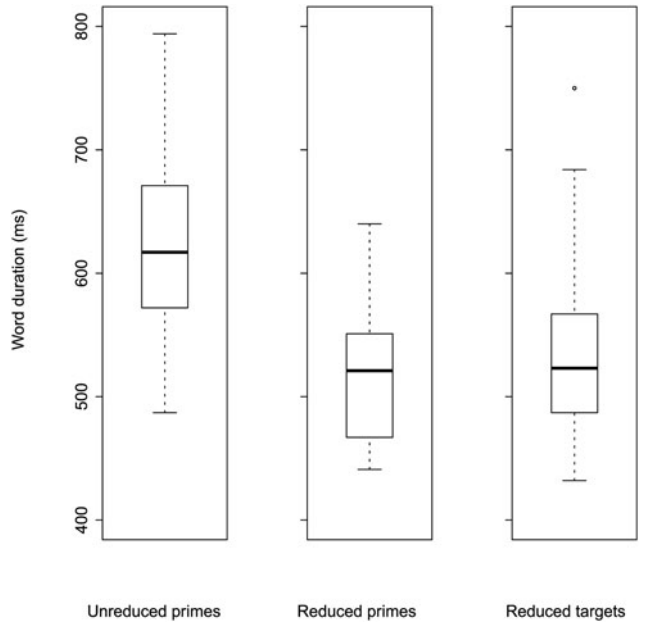


Fig. 3. Distribution of the durations (in ms) of the unreduced primes (left panel), the reduced primes (middle panel) and the reduced targets (right panel) in Experiment 2.

our analyses, we excluded targets whose primes received incorrect responses (8% of the data; 61 targets evaluated by the native English listeners, 96 by the Dutch listeners, and 171 by the Spanish listeners). After this step, listeners overall made 4% of errors on average (SD = 4%; native English speakers: 3%; Dutch listeners: 4%; Spanish listeners: 6%; SD = 3%; 4%; 4%; respectively).

The errors showed no effects of the control variables. More importantly, there was an effect of Listener group, indicating that the English natives made significantly fewer mistakes than the Dutch and Spanish listeners (Contrast 1; $\beta = 0.63$, $SE = 0.27$, $z = 2.33$, $p = 0.019$ in the final model). There was no significant difference between the Dutch and the Spanish listeners (Contrast 2; $\beta = 0.11$, $SE = 0.34$, $z = 0.32$, $p = .75$, see table A2-3). Variant match did not show a simple effect either ($\beta = 0.088$, $SE = 0.21$, $z = 0.42$, $p = 0.67$), and both interactions appeared not significant either ($\beta_{\text{variant match x contrast 1}} = 0.73$; $SE = 0.49$, $z = 1.50$, $p = .13$; $\beta_{\text{variant match x contrast 2}} = 0.70$; $SE = 0.45$, $z = 1.55$, $p = .12$) (again see table A2-3). As noted in the Method section, the final model was obtained by systematically simplifying the full model.

RTs from word onset, after excluding targets with incorrect responses (which accounted, together with the targets whose primes received incorrect responses, for 14% of the data; consisting of 87 targets evaluated by the English native listeners, 138 by the Dutch listeners, and 222 by the Spanish listeners) as well as outliers relative to the grand mean (accounting for 3% of the remaining data; 7 targets evaluated by the English native listeners, 34 by the Dutch listeners, and 34 by the Spanish listeners) ranged from 536 ms to 1679 ms, and were 917 ms on average (SD = 205 ms).

Table A2-4 in Appendix 2 shows the full RT model with the coefficients of both statistically significant and non-significant predictors.

Our final statistical model for the log RTs, summarized in Table 2, shows that Log RT prime and the Log RT on the previous trial were highly significant predictors for log RTs (log RT prime: β=0.16, SE = 0.014, t = 11.47, p = 1.0 e-16; log RT previous trial: β = 0.064; SE = 0.011, T = 5.88, p = 4.1 e-9), indicating that responses were significantly quicker to targets whose primes and whose preceding trials received quicker responses.

More importantly for our research question, we obtained statistically significant interactions between Variant match and Listener group for both contrasts. Figure 4 illustrates the Variant match effects for all listener groups.

To interpret the first Listener group x Variant match interaction, we ran models with the same predictors that were present in the final model for all listeners, apart from the simple effect of Listener group and its interaction term, on the individual subsets of the native listener data and of non-native listener data (to interpret Contrast 1). In these models, we found significant exemplar effects for the non-native listeners ($\beta_{\text{variant match}} = -0.021$, $SE = 0.0070$, $t = -3.01$), but not for the native listeners ($\beta_{\text{variant match}} = 0.012$, $SE = 0.0085$, $t = 1.41$).

In addition, we ran models on the separate data of the Dutch and the Spanish listeners (to interpret Contrast 2). In these models, we observed significant exemplar effects for the Spanish listeners ($\beta_{\text{variant match}} = -0.034$, $SE = 0.011$, $t = -3.11$), but not for the Dutch listeners ($\beta_{\text{variant match}} = -0.0086$, $SE = 0.0086$, $t = -0.99$). All these models on separated data sets are available in the Supplementary Materials (Supplementary Materials).

General discussion

The present study compares the importance of exemplars in native and non-native listening. Our first research question concerns whether there are differences in exemplar effects between native and non-native word comprehension (rather than in old-new judgments or word repetition). Our second research question concerns whether potential differences in exemplar effects between native and non-native listeners depend on the listener's familiarity with the variation pattern from their native language, which may affect the available processing resources.

We investigated these questions with two long-term priming experiments. In Experiment 1 we studied exemplar effects using a language-general variation type (speaker voice), while in Experiment 2 we studied exemplar effects by using a language-specific variation type (vowel reduction in English). In both experiments, we tested native English listeners as well as Dutch and Spanish non-native listeners of English. While Dutch listeners are familiar from their native language with the language-specific variation type studied in Experiment 2, Spanish listeners are not.

While both groups of non-native listeners were administratively determined to be at CEFR level B2 in English, in both experiments, the Dutch listeners obtained slightly higher scores than the Spanish listeners on our proficiency test. In analyses on subsets of the non-native listeners with similar proficiencies, we found the same results as in our analyses performed on all listeners.

In both experiments, participants performed lexical decision. In the first parts of the experiments, they heard all experimental real words for the first time, mixed with pseudowords. In the second parts of the experiments, all real and pseudowords from the first parts reoccurred, mixed with new real words and pseudowords (such that the percentage of reoccurring words was 75%). Theoretically, participants could have performed the lexical decision task for the reoccurring words in the second parts of the experiments without lexical access: they could have adopted a strategy according to which they first determined for each word

https://doi.org/10.1017/S1366728922000116 Published online by Cambridge University Press

Table 2. Final statistical model for log RTs of correct responses to targets in Experiment 2. The intercept represents *Variant mismatch*. *SE* stands for Standard Error. Contrast 1 compares native (coefficient 0.666) to non-native listeners (both -0.333), and Contrast 2 compares Dutch (0.5) to Spanish (-0.5) non-native listeners.

Fixed effects	*B*	*SE*	*t*	*p*
Intercept	5.21	0.12	42.7	0.0
Variant match (match)	−0.011	0.0055	−2.02	0.043
Log RT prime	0.164	0.014	11.47	1 e-16
Log RT previous	0.064	0.011	5.88	4.1 e-9
Contrast 1 (native vs. non-native)	−0.876	0.019	−4.46	1.1 e-5
Contrast 2 (Dutch vs. Spanish)	−0.091	0.022	−4.21	2.6 e-5
Variant match (match) x Contrast 1	0.0369	0.012	3.18	0.0015
Variant match (match) x Contrast 2	0.030	0.013	2.20	0.027
Random effects			*SD*	
Word	Intercept		0.052	
Listener	Intercept		0.084	
Residual			0.139	

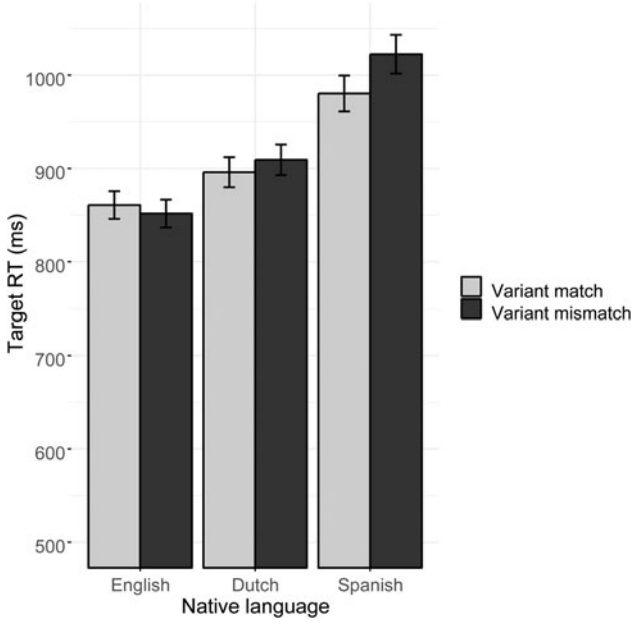


Fig. 4. (Raw, i.e., non log-transformed) RTs of correct responses to targets in Experiment 2, split by listeners' native languages and variant match condition. Error bars represent 95% confidence intervals.

whether they had heard the word before, and if so, in the second step responded with the same response they had given before, and otherwise determined their decision on lexical access. We believe, however, that it is unlikely that participants adopted this complex strategy, for several reasons. First, the two parts of the experiments directly followed each other, without any indication when the second parts started. It would have taken some time before the participants realized that a sufficient number of words reoccurred to make this strategy work. Secondly, this strategy implies that participants not only had to remember which words reoccurred but also which decisions they had taken for these words, making it a complex memory task, which is not reflected in the error scores. Finally, this strategy may seriously delay the decisions for the non-occurring words, because participants would only start lexical access for these words, after having established that these words had not occurred before. We do not find evidence for this delay[1]. We thus assume that participants performed lexical decision via lexical access throughout the complete experiments.

In Experiment 1, where we manipulated speaker voice, we only found clear exemplar effects for native listeners. The difference in exemplar effects between native and non-native listeners speaks to our first research question. Our study is the first to clearly establish this difference for word comprehension, as the previous studies that compared native and non-native listening did not report or find statistically significant differences, and/or tested participants in tasks that do not directly require word comprehension (Drozdova et al., 2019; Trofimovich, 2005; Winters et al., 2013). Our finding may be indicative of qualitative differences between native and non-native speech comprehension, which has consequences for models of L2 speech comprehension (e.g., the 'bilingual interactive activation plus' (BIA+) model developed by Dijkstra & Heuven, 2002).

We propose that the difference in the presence versus absence of exemplar effects observed between native and non-native listeners in Experiment 1 may be accounted for by the availability of more processing resources for native than for non-native listeners. While lexical decision is quite an easy task for native listeners (see the many plain lexical decision experiments where native listeners produce hardly any errors, e.g., Brand & Ernestus, 2018), this is not true for non-native listeners. Non-natives' weaker and less well-specified lexical representations (e.g., Cook et al., 2016) introduce a considerable amount of uncertainty to the task, which makes it harder for non-native listeners. This explanation for the absence of exemplar effects for the non-native

[1]We examined whether the reaction times on words and pseudo words in the second part of Experiment 1 that had not occurred before were longer than the reaction times to words and pseudo words presented in the first part of the Experiment. This was not the case. Also for Experiment 2 it was verified that there was no such statistically significant delay. All (insignificant) differences in RTs between primes in part 1 and non-recurring targets in part 2 could be explained by Trial number. For both Experiments 1 and 2, the corresponding lmer models are provided in the Supplementary Materials.

https://doi.org/10.1017/S1366728922000116 Published online by Cambridge University Press

listeners is in line with previous work with native listeners which showed that moderately challenging conditions may lead to smaller exemplar effects (e.g., McLennan et al., 2003; Nygaard et al., 2000).

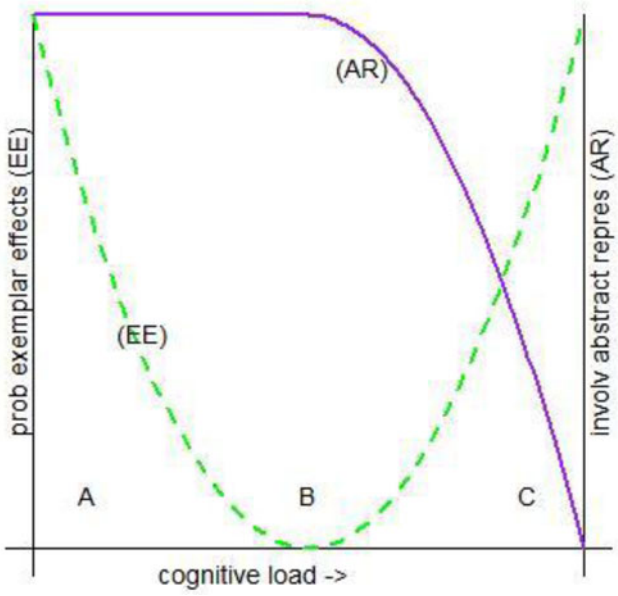
This explanation is also in line with the fact that studies reporting exemplar effects for speaker voice for non-native listeners used tasks requiring less processing effort than lexical decision. In Drozdova et al. (2019) and Winters et al. (2013), exemplar effects arose in an old-new judgment task, which does not involve a costly lexical search like lexical decision does. In fact, one of the listener groups in the Winters et al. study was able to perform the old-new judgment task without knowledge of the target language, again showing that this task could be performed with little to no lexical involvement. The two experiments that failed to show exemplar effects for non-native listeners involve lexical access (word identification in Drozdova et al., 2019) or another task involving high processing load (word repetition, Trofimovich, 2005). Taken together, these results strongly suggest that processing costs (which can differ between native and non-native listeners and by task) can affect the emergence of exemplar effects.

In Experiment 2, primes and targets were produced by the same speaker, but matched or mismatched in their degree of speech reduction. Reduced word tokens had highly shortened vowels in their initial syllables. This reduction occurs frequently in English (e.g., Shockey, 2003) and Dutch (Ernestus, 2000), mostly in words embedded in sentences. In contrast, Spanish does not reduce its vowels as much. This difference in familiarity may lead to a difference in processing cost (cf. Adank, Evans, Stuart-Smith & Scotti, 2009) between Dutch and Spanish non-native listeners of English, which, in turn, may lead to a difference between these listener groups in the emergence of exemplar effects arising from vowel reduction.

In line with this hypothesis, our results in Experiment 2 showed that only the Spanish listeners relied on exemplars. This finding is in line with results reported by Morano et al. (2019). They also showed that exemplar effects may arise for non-native listeners in lexical decision when these listeners are confronted with variation that they are not very familiar with from their L1.

We believe that the combination of the experimental outcomes from Experiments 1 and 2 cannot be completely reconciled in any existing theoretical framework of auditory word processing. We tentatively propose that, like the results from Experiment 1, those from Experiment 2 can also be explained by the role of processing load on auditory word recognition. We suggest that the exact patterns observed in our experimental data from Experiments 1 and 2 may be explained as a function of processing load as follows (see also Figure 5). If listeners are familiar with a particular type of variation (as all listener groups in Experiment 1), exemplar effects can be expected (condition A in Figure 5), unless the experimental task requires high processing load, such as lexical decision for non-native listeners. In case of high processing load, the use of abstract representations is most efficient in lexical decision tasks, which explains why the non-native listeners in Experiment 1 do not show exemplar effects (condition B in Figure 5).

Also for native listeners, processing load can be higher than usual – for example, if the variation they hear occurs in an unfamiliar condition (such as reduction in words in isolation; see, e.g., Brand & Ernestus, 2018, where native listeners of French almost flawlessly responded to unreduced French words in a lexical decision task, while their accuracy was lower than 90% for the same words when the words' initial schwa was reduced). This may explain why the native listeners do not show exemplar effects in Experiment 2 (condition B in Figure 5). The role of processing load thus explains why the native listeners behaved differently in Experiments 1 and 2: In Experiment 1, variation was familiar and the lexical decision task did not involve high processing costs, yielding exemplar effects. In Experiment 2, processing costs were higher due to the occurrence of reduced words in isolation, leading to the use of abstract representations, without exemplar effects. Similarly, the role of processing load explains why the natives in Experiment 2 behaved like the non-native listeners in Experiment 1: both groups faced moderate processing costs that made them ignore exemplars.



The listener group in this study with the highest processing load is the Spanish in Experiment 2, who were confronted with both a lexical decision task and with variation that was highly unfamiliar to them from their L1. We assume that they could not efficiently process the variation, while they noticed that many words were repeated. This may have invited them to focus on the detailed acoustic properties of the words, which yielded exemplar effects (condition C in Figure 5). This reliance on exemplars may offer a processing benefit in a context where difficult words are repeated. For the Dutch listeners, the type of reduction used in Experiment 2 was familiar from their L1, and their processing load was therefore considerably smaller than the Spanish listeners' load. They could therefore still make use of abstract representations, with no exemplar effects as a result (condition B in Fig. 5).

In sum, we propose that processing load determines whether listeners rely on both lexical representations and exemplars, or especially on one of these representations. If processing load is low, exemplar effects may emerge if additional specific conditions are met (e.g., lags between the two occurrences of a word are not too large, Hanique et al., 2013). If processing load is moderate to high, listeners in general rely on an abstract representation (but in

https://doi.org/10.1017/S1366728922000116 Published online by Cambridge University Press

specific cases episodic effects might show up, McLennan & Luce, 2005). If the processing load is extremely high, listeners will mostly rely on exemplars. We welcome new studies testing this proposed relation between exemplar effects and processing load.

In both experiments, we tested for exemplar effects in one direction: in Experiment 1, a match between prime and target was always implemented as male (prime) – male (target), and a mismatch as female (prime) – male (target); all targets were thus produced by the same male speaker. In Experiment 2, a prime-target match was always implemented as reduced (prime) – reduced (target), and a mismatch as unreduced (prime) – reduced (target). Here, all targets were reduced tokens. Because of this set-up, it is unclear whether the exemplar effects obtained in Experiment 1 may generalize to targets produced by the female speaker, and in Experiment 2 to unreduced targets. An experimental design testing for exemplar effects in both directions would be necessary to establish this. Future experiments should directly test these relations in more detail.

In conclusion, we documented differences in the conditions under which native and non-native listeners may show exemplar effects in auditory identity priming experiments. Our results are in line with our tentative hypothesis that the emergence of exemplar effects is modulated by the available processing resources, which, in their turn, are modulated by listeners' familiarity with the variation patterns. Because native and non-native listeners may differ in their processing load, they would consequently differ in their reliance on exemplars. This suggests that the type of representations (abstract representations or exemplars) primarily used in speech comprehension may differ between native and non-native listening, which points at qualitative differences between native and non-native listening.

Acknowledgements. This work was supported by a VICI grant (277-70-010) from the Netherlands Organization for Scientific Research (NWO) awarded to the third author. This work was carried out while the first author was affiliated at Radboud University, Nijmegen, the Netherlands.

Supplementary Material. For supplementary material accompanying this paper, visit http://dx.doi.org/10.1017/S1366728922000116

References

Adank P, Evans B, Stuart-Smith J and Scotti S (2009) Comprehension of familiar and unfamiliar native accents under adverse listening conditions. *Journal of Experimental Psychology: Human Perception and Performance*, **35** (2). pp. 520–529. ISSN 0096-1523

Baayen RH (2008) *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Barr DJ, Levy R, Scheepers C and Tily HJ (2013) Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* **68**, 255–278. doi: 10.1016/j.jml.2012.11.001

Bates D, Kliegl R, Vasishth S and Baayen H (2015) Parsimonious Mixed Models. *arXiv*. 1506.

Bates D, Maechler M, Bolker B and Walker S (2015) Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, **67**, 1–48. doi: 10.18637/jss.v067.i01

Boersma P and Weenink D (2018) Praat: doing phonetics by computer (Version 6.0.43).

Brand S and Ernestus M (2018) Listeners' processing of a given reduced word pronunciation variant directly reflects their exposure to this variant: evidence from native listeners and learners of French. *Quarterly Journal of Experimental Psychology* **71**, 1240–1259. doi:10.1080/17470218.2017.1313282

Church BA and Schacter DL (1994) Perceptual specificity of auditory priming: implicit memory for voice intonation and fundamental frequency. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **20**, 521–533. doi: 10.1037/0278-7393.20.3.521

Cook SV, Pandža NB, Lancaster AK and Gor K (2016) Fuzzy nonnative phonolexical representations lead to fuzzy form-to-meaning mappings *Frontiers in Psychology* **7**, 1–17. doi: 10.3389/fpsyg.2016.01345

Craik F. I. M. and Kirsner K (1974) The effect of speaker's voice on word recognition. *Quarterly Journal of Experimental Psychology* **26**, 274–284. doi: 10.1080/14640747408400413

Dalby J (1986) Phonetic structure of fast speech in American English. Ph.D. dissertation, Indiana University.

Dijkstra T and Heuven W. J. B. V. (2002) The architecture of the bilingual word recognition system: From identification to decision. *Bilingualism: Language and Cognition*. **5**, 175–197. doi: 10.1017/S1366728902003012

Drozdova P, van Hout RV and Scharenborg O (2019) Talker-familiarity benefit in non-native recognition memory and word identification: The role of listening conditions and proficiency. *Attention, Perception, and Psychophysics* **81**, 1675–1697. doi: 103758/s13414-018-01657-5

Edstrom A (2006) L1 use in the L2 classroom: One teacher's self-evaluation. *The Canadian Modern Language Review* **63**, 275–292.

Ernestus M (2000) Voice assimilation and segment reduction in casual Dutch: A corpus-based study of the phonology-phonetics interface. Ph.D. dissertation, Utrecht, Nederland: LOT.

Felker E, Ernestus M and Broersma M (2019) Evaluating dictation task measures for the study of speech perception. In Calhoun S, Escudero P, Tabain M and Warren P (eds), *Proceedings of the 19th International Congress of Phonetic Sciences, Melbourne, Australia 2019*. Canberra, Australia: Australasian Speech Science and Technology Association Inc, pp. 383–387. doi: 10.1037/0022-0663.99.1.154

Fox J and Monette G (2002) *R and S-Plus Companion to Applied Regression*. Thousand Oaks, CA, USA: Sage Publications, Inc.

Goh WD (2005) Talker variability and recognition memory: instance-specific and voice-specific effects. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **31**, 40–53. doi: 10.1037/02787393.31.1.40

Goldinger SD (1996) Words and voices: episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **22**, 1166–83. doi: 10.1037/0278-7393.22.5.1166

Goldinger SD (2007) A complementary-systems approach to abstract and episodic speech perception. *Proceedings of the 16th International Congress of Phonetic Sciences*.

Hanique I, Aalders E and Ernestus M (2013) How robust are exemplar effects in word comprehension? *Mental Lexicon* **8**, 269–294. doi: 10.1075/ml.8.3.01han

Janse E (2008) Spoken-word processing in aphasia: effects of item overlap and item repetition. *Brain and Language* **105**, 185–198. doi: 10.1016/j.bandl.2007.10.002

Jones K and Ono T (2000) Reconciling textbook dialogues and naturally occurring talk: What we think we do is not what we do. *Arizona Working Papers in SLAT*.

Krestar ML and McLennan CT (2013) Examining the effects of variation in emotional tone of voice on spoken word recognition. *Quarterly Journal of Experimental Psychology* **66**, 1793–1802. doi: 10.1080/17470218.2013.766897

Lemhöfer K and Broersma M (2012) Introducing LexTALE: a quick and valid Lexical Test for Advanced Learners of English. *Behavior Research Methods* **44**, 325–343. doi: 10.3758/s13428-011-0146-0

Lo S and Andrews S (2015) To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology* **6** doi: 10.3389/fpsyg.2015.01171

McCarthy M and Carter R (1995) Spoken grammar: What is it and how can we teach it? *ELT Journal* **49**, 207–218.

McLennan CT and Luce PA (2005) Examining the time course of indexical specificity effects in spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **31**, 306–321. doi: 10.1037/0278-7393.31.2.306

McLennan CT, Luce PA and Charles-Luce J (2003) Representation of lexical form. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **29**, 539–553. doi: 10.1037/0278-7393.29.4.539

Mitterer H and Tuinman A (2012) The role of native-language knowledge in the perception of casual speech in a second language. *Frontiers in Psychology* **3**. doi: 10.3389/fpsyg.2012.00249

Morano L, ten Bosch L and Ernestus M (2019) Looking for exemplar effects: testing the comprehension and memory representations of r'duced words in Dutch learners of French. In Fuchs S, Rochet-Capella A and Cleland J (eds), *Speech Perception and Production: Learning and Memory*. Bern, pp. 1–29.

Nijveld A, ten Bosch L and Ernestus M (2015) Exemplar effects arise in a lexical decision task, but only under adverse listening conditions. Exemplar effects arise in a lexical decision task, but only under adverse listening conditions. In Wolters M, Livingstone J, Beattie B, Smith R, MacMahon M, *et al.* (eds), *Scottish consortium for ICPhS, Proceedings of the 18th International Congress of Phonetic Sciences (ICPhS 2015)*. Glasgow: University of Glasgow.

Nygaard LC, Burt SA and Queen JS (2000) Surface form typicality and asymmetric transfer in episodic memory for spoken words. *J Exp Psychol Learn Mem Cogn* **26**(5), pp. 1228–1244. doi: 10.1037//0278-7393.26.5.1228.

Palmeri TJ, Goldinger SD and Pisoni DB (1993) Episodic encoding of voice attributes and recognition memory for spoken words. *Journal of Experimental Psychology: Learning, Memory, and Cognition* **19**, 309–328. doi: 10.1037/0278-7393.19.2.309

Pufahl A and Samuel AG (2014) How lexical is the lexicon? Evidence for integrated auditory memory representations. *Cognitive Psychology* **70**, 1–30. doi: 10.1016/j.cogpsych.2014.01.001

R Core Team (2017) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/

Segalowitz N (2010) *Cognitive bases of second language fluency*. Routledge, 2010.

Shockey L (2003) *Sound patterns of spoken English*. Blackwell. doi: 10.1002/9780470758397

Sumner M and Samuel AG (2005) Perception and representation of regular variation: The case of final /t/. *Journal of Memory and Language* **52**, 322–338. doi: 10.1016/j.jml.2004.11.004

Ten Bosch L, Giezenaar G, Boves L and Ernestus M (2016) Modeling language learners' errors in understanding casual speech. In Adda G, Barbu Mititelu V, Mariani J, Tufiş D and Vasilescu I (eds), *Errors by humans and machines in multimedia, multimodal, multilingual data processing. Proceedings of Errare 2015*. Bucharest: Editura Academiei Române, pp. 107–121

Torreira T and Ernestus M (2011) Realization of voiceless stops and vowels in conversational French and Spanish. *Laboratory Phonology* **2**, 331–353. doi:10.1515/labphon.2011.012.

Trofimovich P (2005) Spoken-word processing in native and second languages: An investigation of auditory word priming. *Applied Psycholinguistics* **26**, 479–504. doi: 10.1017/S0142716405050265

Wilder RJ (2018) Investigating hybrid models of speech perception. Dissertation 2018, University of Pennsylvania.

Winters S, Lichtman K and Weber S (2013) The role of linguistic knowledge in the encoding of words and voices in memory. In Voss E, Tai SD and Li Z (eds), *Selected Proceedings of the 2011 Second Language Research Forum: Converging Theory and Practice*. Somerville, MA, USA: Cascadilla Proceedings Project, pp. 129–138

https://doi.org/10.1017/S1366728922000116 Published online by Cambridge University Press

Appendix 1: Stimulus materials

Table A1-1. Stimuli occurring in the experiments (excluding the three practice items), with raw (Freq.) and log-transformed (log-freq.) frequencies of occurrence for the real words. The repeated real words are the experimental real words.

Repeated stimuli				Non-repeated stimuli			
Real word	Freq.	Log-freq.	Pseudo word	Real word	Freq.	Log-freq.	Pseudo word
balloon	6	2.68	ballee	recall	6	2.63	rekel
banana	5	2.39	benooga	research	253	7.98	resers
belief	51	5.66	beleesh	repair	15	3.88	repor
career	75	6.23	kerame	vanilla	2	0.83	vanole
cassette	7	2.90	kaseet	supply	69	6.11	supplee
cement	6	2.57	semont	deposit	17	4.09	depaset
collapse	21	4.37	coliss	canoe	4	1.96	canee
committee	189	7.56	komanee	result	186	7.54	rezell
debate	69	6.11	debome	selection	60	5.91	selaksin
decline	41	5.35	decloof	salute	2	1.20	saluke
defeat	29	4.86	defoose				
defect	6	2.68	defess				
defence	115	6.84	defots				
degree	98	6.62	degoo				
delay	23	4.50	delow				
design	119	6.89	dezone				
disease	89	6.47	dezoom				
divorce	17	4.05	devees				
domain	17	4.05	domoon				
guitar	27	4.75	guitee				
machine	82	6.36	mechoon				
parade	10	3.33	parogue				
police	270	8.08	poloose				
potato	9	3.10	potono				
safari	2	1.19	saforro				
salami	1	0.19	saleemo				
saloon	5	2.44	saleen				
surprise	45	5.48	suppees				
tobacco	15	3.87	tabodo				
tomato	7	2.85	tomeeno				

https://doi.org/10.1017/S1366728922000116 Published online by Cambridge University Press

Appendix 2: Other statistical models

In Appendix 2, we list four statistical models. Two models are presented for the accuracy of the correct responses to targets for both experiments. In line with recent multidisciplinary discussions about p-values, the statistical package R does not provide them by default in (generalized) linear mixed effect models, but for convenience we have added them in the tables below. In addition, two corresponding full models are presented for the log RT of correct responses to targets in both experiments. All four models are full models in the sense that they include all fixed effects we tested (whether or not they were significant) that did not lead to model convergence issues.

Table A2-1. Statistical model for the accuracy of responses to targets in Experiment 1 including all fixed predictors that did not lead to model convergence issues (whether or not statistically significant). *Speaker mismatch* is on the intercept (the same model with *Speaker match* on the intercept yielded convergence issues), *SE* stands for Standard Error. Contrast 1 compares native (0.666) to non-native listeners (both -0.333), and Contrast 2 compares Dutch (0.5) to Spanish (-0.5) listeners. We did not include Listener group as random slope by Word because inclusion of this slope produced model convergence issues, and we did not include Speaker match as random slope by Word or by Listener because inclusion of each of these slopes led to singular fits. Further inclusion of predictors lag, trial number or word duration yielded divergent models and so are discarded.

Fixed effects	*B*	*SE*	*z*	*p*
Intercept	3.62	0.48	7.58	3.4 e-14
Speaker match (match)	0.37	0.22	1.69	0.091
Log target frequency	0.0016	0.09	0.018	0.99
Contrast 1 (native vs. non-native)	0.32	0.33	0.97	.33
Contrast 2 (Dutch vs. Spanish)	0.27	0.38	0.71	.48
Variant match x Contrast 1 (native vs. non-native)	0.26	0.48	0.54	.59
Variant match x Contrast 2 (Dutch vs. Spanish)	0.10	0.53	0.20	.84
Random effects				*SD*
Word	Intercept			0.68
Listener	Intercept			0.61

Table A2-2. Statistical model for log RTs of correct responses (from the English, Spanish and Dutch participants) to targets in Experiment 1 displaying the effects of all fixed predictors (whether statistically significant or not). The intercept represents *Speaker*. *SE* stands for Standard Error. Contrast 1 compares native (0.666) to non-native listeners (both -0.333), and Contrast 2 compares Dutch (0.5) to Spanish (-0.5) non-native listeners. Random effects were included only insofar they yielded convergent models.

Fixed effects	*B*	*SE*	*t*	*p*
Intercept	4.18	0.48	8.68	1e-16
Speaker match (match)	−0.0039	0.052	−0.74	0.46
Log RT previous	0.062	0.011	5.52	3.8 e-8
Log RT prime	0.180	0.014	12.58	1 e-16
Log target duration	0.150	0.073	2.04	0.041
Log target frequency	−0.00038	0.0045	−0.084	0.93
Trial	0.00018	0.00017	0.95	0.34
Lag	−0.00011	0.00017	−0.64	0.52
Contrast 1 (native vs. non-native)	−0.067	0.016	−4.20	2.6 e-5
Contrast 2 (Dutch vs. Spanish)	−0.0049	0.019	−0.25	0.80
Speaker match (match) x Contrast 1	−0.023	0.011	−2.19	0.029
Speaker match (match) x Contrast 2	−0.011	0.013	−0.89	0.37
Random effects			*SD*	
Word	Intercept		0.044	
Listener	Intercept		0.069	
Residual			0.132	

https://doi.org/10.1017/S1366728922000116 Published online by Cambridge University Press

Table A2-3. Statistical model predicting the accuracy of targets in Experiment 2 displaying effects of all fixed predictors that did not lead to model convergence issues (whether or not statistically significant). We present the full model with the interaction between Variant match and Listener group, with Trial as additional predictor. The predictors lag, word duration, and word frequency led to divergent models and so are not included here.
Variant mismatch is on the intercept, *SE* stands for Standard Error. Contrast 1 compares native (0.666) to non-native listeners (both -0.333), and Contrast 2 compares Dutch (0.5) to Spanish (-0.5) non-native listeners. We did not include random slopes in our final model because their inclusions either led to model convergence issues(Variant match by Word) or to a singular fit (Listener group by Word and Variant match by Listener).

Fixed effects	*B*	*SE*	*z*	*p*
Intercept	4.47	0.68	6.51	7.7 e-11
Variant match (match)	0.09	0.21	0.42	0.67
Contrast 1 (native vs. non-native)	0.31	0.33	0.96	0.34
Contrast 2 (Dutch vs. Spanish)	0.11	0.34	0.32	.0.74
Trial	−0.006	0.0055	−1.13	0.26
Variant match x Contrast 2 (Dutch vs. Spanish)	0.73	0.49	1.50	0.13
Variant match x Contrast 2 (Dutch vs. Spanish)	0.70	0.45	1.55	.0.12
Random effects				*SD*
Word	Intercept			1.22
Listener	Intercept			0.63

Table A2-4. Statistical model for log RTs of correct responses to targets in Experiment 2 displaying effects of all fixed predictors (whether or not statistically significant). *Variant mismatch* is on the intercept. *SE* stands for Standard Error. Contrast 1 compares native (0.666) to non-native listeners (both -0.333), and Contrast 2 compares Dutch (0.5) to Spanish (-0.5) non-native listeners. We did not include Listener group, Variant match or their interaction as random slopes by Word because their inclusions each led to model convergence issues, and we did not include Variant match as slope by Listener because of a perfect correlation with the intercept ($r = 1.0$).

Fixed effects	*B*	*SE*	*T*	*p*
Intercept	4.77	0.51	9.35	1 e-16
Variant match (match)	−0.010	0.0055	−1.82	0.068
Log RT previous	0.070	0.011	6.41	1.4 e-10
Log RT prime	0.16	0.014	11.39	1.0 e-16
Log target duration	0.067	0.078	0.86	0.39
Log target frequency	−0.0017	0.0054	−0.31	0.76
Trial	0.00016	0.00021	0.80	0.42
Lag	−0.00034	0.00018	−1.85	0.064
Contrast 1 (native vs. non-native)	−0.085	0.019	−4.38	1.7 e-5
Contrast 2 (Dutch vs. Spanish)	−0.089	0.022	−4.12	3.7 e-5
Variant match (match) x Contrast 1	0.035	0.012	3.01	0.026
Variant match (match) x Contrast 2	0.027	0.013	2.05	0.04
Random effects			*SD*	
Word	Intercept		0.052	
Listener	Intercept		0.085	
Residual			0.14	

https://doi.org/10.1017/S1366728922000116 Published online by Cambridge University Press